OPEN ACCESS

University of BRISTOL

Rodriguez, S., Alghamdi, O., Guthrie, P., Shihab, H., McArdle, W., Gaunt, T., Alharbi, K. K., & Day, I. (2017). Frequency of *KLK3* gene deletions in the general population. *Annals of Clinical Biochemistry*, *54*(4), 472-480. https://doi.org/10.1177/0004563216666999

Peer reviewed version

Link to published version (if available):
10.1177/0004563216666999

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Sage at http://journals.sagepub.com/doi/10.1177/0004563216666999. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

Frequency of *KLK3* gene deletions in the general population

Short title: Frequency of *KLK3* deletions

Santiago Rodriguez[1†*], Osama A. Al-Ghamdi[2,3†], Philip A. I. Guthrie[1], Hashem A. Shihab[1], Wendy McArdle[3], Tom Gaunt[1], Khalid K. Alharbi[2], and Ian N. M. Day[3]

[1] MRC Integrative Epidemiology Unit (IEU), School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK

[2] Clinical Laboratory Sciences Department, College of Applied Medical Sciences, King Saud University, P. O. Box 10219, Riyadh 11433, Saudi Arabia

[3] School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK

[†] Equal first authorship

[*] To whom correspondence should be addressed: Santiago Rodriguez, MRC Integrative Epidemiology Unit (IEU), School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, United Kingdom. Tel: +44 01173310133; Email: santi.rodriguez@bristol.ac.uk

Word count: 2850

COMPETING INTERESTS

The authors declare that they have no competing interests.

FUNDING

This work was supported by an overseas PhD studentship from King Saud University (Riyadh, Saudi Arabia) to Osama Al-Ghamdi. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. The Integrative Epidemiology Unit is supported by the MRC and the University of Bristol (MC_UU_12013/1-9).

ETHICAL APPROVAL

Ethical approval was obtained from the South East, UK, Multi-centre Research Ethics Committee (reference 01/1/44).

ACKNOWLEDGEMENTS

This work was supported by an overseas PhD studentship from King Saud University (Riyadh, Saudi Arabia) to Osama Al-Ghamdi. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. The Integrative Epidemiology Unit is supported by the MRC and the University of Bristol (MC_UU_12013/1-9).

ABSTRACT

BACKGROUND: One of the kallikrein genes (*KLK3*) encodes Prostate Specific Antigen (PSA), a key biomarker for prostate cancer. A number of factors, both genetic and non-genetic, determine variation of serum PSA concentrations in the population. We have recently found three *KLK3* deletions in individuals with very low PSA levels, suggesting a link between abnormally reduced *KLK3* expression and deletions of *KLK3*. Here we aim to determine the frequency of kallikrein gene 3 deletions in the general population. METHODS: The frequency of *KLK3* deletions in the general population was estimated from the 1958 Birth Cohort sample (N=3815) using ARCS (Amplification Ratiometry Control System). *In silico* analyses using PennCNV were carried out in the same cohort and in NBS-WTCCC2 in order to provide an independent estimation of the frequency of *KLK3* deletions in the general population. RESULTS: ARCS results from the 1958 cohort indicated a frequency of *KLK3* deletions of 0.81% (3.98% following a less stringent calling criterion). From *in silico* analyses, we found that potential deletions harbouring the *KLK3* gene occurred at rates of 2.13% (1958 Cohort, N=2867) and 0.99% (NBS-WTCCC2, N=2737) respectively. These results are in good agreement with our *in vitro* experiments. All deletions found were in heterozygosis. CONCLUSIONS: We conclude that a number of individuals from the general population present *KLK3* deletions in heterozygosis. Further studies are required in order to know if interpretation of low serum PSA concentrations in individuals with *KLK3* deletions may offer false negative assurances with consequences for prostate cancer screening, diagnosis and monitoring.

Keywords: *KLK3*; PSA; ARCS; PennCNV; deletions

INTRODUCTION

Prostate Specific Antigen (PSA) is a key biomarker for prostate cancer (PCa) diagnosis and monitoring, and has been advocated for PCa screening.[1] PSA is encoded by *KLK3*, a kallikrein gene.[2] Several studies have associated serum PSA concentrations with *KLK3* SNPs.[3] Most of the associations in the pre-Genome Wide Association Studies (GWAS) era were insufficiently replicated. However, recent GWAS[4-6] with higher power have reinforced a number of previous associations, in addition to discovering new variants linked with circulating PSA.

We have previously analysed the presence of inactivating genetic variants in *KLK3* that would be consistent with haploinsufficient levels of PSA.[7] A thorough analysis was done to investigate *KLK3* variants at the DNA sequence and at the structural levels from the study of thirty individuals with the lowest PSA concentrations (≤0.1 ng/mL) and absence of PCa among 85,000 men from the ProtecT biorepository.[7] This report was the only study to date that detected the presence of hemizygous deletions in three individuals with very low levels of PSA, representing a likely link between *KLK3* deletions and low PSA concentrations.[7] These individuals had no diagnosis of PCa. However, other (affected) individuals with these genetic rearrangements may be misidentified as PCa-free due to their impaired ability to express PSA.

PCR-based methods have prevailed as the tools of choice for the targeted screening of Copy Number Variants (CNVs), due to their rapidity, simplicity and flexibility. Several techniques employ PCR in a multiplex manner, as a means to enhance gene dosage estimation by co-amplifying reference genes along with ones in test. Examples include multiplex polymerase

chain reaction of short fluorescent fragments,[8] multiplex amplifiable probe hybridisation (MAPH),[9] multiplex ligation-dependent probe amplification (MLPA),[10] the paralogue ratio test (PRT)[11] and quantitative real time PCR (qPCR), in addition to the methodology used in this study: the amplification ratiometry control system (ARCS).[12] Apart from ARCS, all of these methods have been reviewed at length [13] and a useful comparison of performance characteristics across these PCR-based tests for copy number is available.[12]

Gross rearrangements can also be detected by *in silico* approaches. It is known that the agreement between both *in silico* and experimental approaches is important for CNV validation.[14]

SNP arrays were originally designed to genotype SNPs for GWAS; however they have been adapted for structural variant discovery from large datasets. Several generations of dense SNP array platforms have been made available by leading companies in the industry (e.g. Illumina and Affymetrix), motivating the development of several algorithms (e.g. PennCNV) that normalise key metrics like signal probe intensities and minor allele frequencies from various forms of genetic data, to detect copy number variants.

PennCNV employs hidden Markov models to infer the underlying states of copy number while interrogating multiple features, two of which are key in CNV detection. The B Allele Frequency (BAF) value is a measure of the allelic intensity (i.e. proportions of the B allele compared to the A allele), whereas Log R Ratio (LRR) is an indicator of signal intensity. PennCNV has been previously used to detect CNVs in genomic regions.[15; 16]

The characterisation of the frequency of *KLK3* deletions in the general population is relevant to the knowledge of their implications for the use of PSA as a biomarker for the detection of PCa, in addition to monitoring its progression, and guiding treatment. Our aim in this study was to estimate the frequency of *KLK3* deletions in the general population from both *in vitro* and *in silico* data.

MATERIALS AND METHODS

WTCCC2 studies: the 1958BC and the NBS samples

The National Child Development Study (NCDS), otherwise known as the 1958 British birth cohort (1958BC) started as a perinatal mortality and morbidity survey, looking at all births in England, Wales and Scotland in a single week in 1958. This included an original sample of 17,638 births (in addition to a further 920 immigrants born in the same reference week). Cohort members were further followed-up by medical examinations (at 7, 11 and 16 years of age) and interviews (at ages 23, 33 and 42). The first biomedical assessment was conducted between September 2002 and March 2004 by trained nurses from the National Centre for Social Research, who visited the homes of cohort members at age 44-45 years.[17, 18]

Consent was sought to collect blood samples for the extraction and storage of DNA and the creation of immortalised cell lines (via Epstein-Barr viral transformation of peripheral blood lymphocytes) for research purpose. A final set of 8018 EDTA blood samples, in addition to 7526 successfully transformed cell cultures were used to derive two DNA sample series of peripheral blood and cell line origin respectively [http://www2.le.ac.uk/projects/birthcohort/1958bc/available-resources]. These efforts led to the creation of a national research resource through funds from the Wellcome Trust and The Medical Research Council.[18]

A subset of 3000 controls from the 1958BC, in addition to 3000 samples from the UK National Blood Service (NBS) were genotyped using cell line DNA by the WTCCC2 (Wellcome Trust Case Control Consortium, second round) in WTCCC2 project controls study (Study ID EGAS00000000028). The 6000 controls were genotyped on both the Illumina 1.2M and the Affymetrix v6.0 chips [http://www.wtccc.org.uk/ccc2/].

Ethical approval was obtained from the South East, UK, Multi-centre Research Ethics Committee (reference 01/1/44).

ARCS assays to estimate the frequency of *KLK3* deletions in the 1958 cohort

A high-throughput ratiometric method (ARCS) developed and validated in our laboratory,[12] was used to identify *KLK3* deletions. This approach has previously been used to detect *KLK3* deletions in individuals with very low PSA levels.[7] In short, it is based on a PCR protocol which analyses the ratio of copy-number-variable target gene to diploid (2n)-copy reference gene. The target and reference gene amplicons are designed to have melting temperatures separated by a few degrees to enable differentiation in a melt assay, and the ratio of target copy number to reference is derived from the change in fluorescence contributed by each as it undergoes melting.

Rodriguez et al,[7] showed that *KLK3* exon 3 is a good indicator of *KLK3* deletions, since it was deleted in all three individuals showing *KLK3* deletions. Copy number status was surveyed for a total of 3815 cell line DNA samples from the 1958 cohort by ARCS at *KLK3* exon 3. Beta globin (*HBB*), an established reference for ARCS assays in our lab, was used as

the reference for diploid number. The assays were run in duplicates for each sample. ARCS primers and conditions were as previously described.[7]

Statistical analysis of data from ARCS

An in-house Perl script written by Guthrie et al[12] was used to calculate peak heights. This script analyses melt files (.mlt) from the LightTyper (Roche Diagnostics Corporation, Indianapolis, Indiana, USA), generating cluster plots to enable the visualisation of potential copy number classes. Further output lists well position, left (L) peak height, right (R) peak height, (L/R) peak height ratio and fluorescence intensity per well in spreadsheet format. These were saved for further statistical analyses. 95% confidence intervals were calculated to assign copy number, as follows: In ARCS, target $_{\text{peak height}}$/reference $_{\text{peak height}}$ ratios are calculated to infer copy number, after which the mean ARCS ratio (a plate specific attribute) for all of the samples in a run is calculated.

In the original implementation of ARCS, the left peak was for a target gene (*HP)*, while the right peak was for a reference amplicon (*TP53).* Deletions or duplications will have ratios lower or higher than 1.96SDs from the mean ratio –respectively-. Whereas, in the *KLK3* assays described above, this order was flipped, assigning the left peak for the AT-rich reference amplicons *HBB*, whereas the right peak represents the target (*KLK3* exon 3) (higher in GC content). This would make the height of the target peak the denominator in calculated ratios, and thus renders it inversely proportional to the ARCS ratios (i.e. deletions will have higher ratios than duplications).

Samples with values that lie outside the ±1.96SDs of the mean target $_{\text{peak height}}$/reference $_{\text{peak height}}$ are flagged as potential CNVs. ARCS assays were designed to test *KLK3* exons 3 (conditions in supplementary file).

PennCNV software to estimate the frequency of *KLK3* deletions in the 1958 and NBS cohorts

We used the Illumina dataset (ega Dataset ID EGAD00000000022). All of the data from iDAT files corresponding to ≈3000 samples from each cohort were compacted into 13 final reports per dataset. These reports contained all of the essential probe intensity data needed by PennCNV.[19] PennCNV, initially built to analyse Illumina genotyping data is capable of handling Affymetrix data through its alternative version PennCNV-Affy. In general, for an algorithm developed to analyse data from a specific platform, its performance is more powerful when applied to data from the same format and shows some weakness when applied to another platform.[14] In addition, algorithm-specific platforms are more specific than platform-independent software.[20]

Individual exclusions criteria

A comprehensive set of genotypes, intensities and signal data were provided for the 1958BC and the NBS controls datasets. Individual exclusions from the initial 3000 samples in each set were based on some of the following considerations: (i) Based on a principal component analysis (PCA) of HapMap individuals, samples that differed from the majority of the collection in terms of ancestry were excluded. The final datasets within the WTCCC2 project

controls study included genotypes from the 1958BC (n=2867) and NBS (n=2737) after initially drawing 3000 DNA samples in each of the two cohorts. (ii) Gender discrepancies between sex reported by suppliers and sex estimated from normalising A-allele probes on the X chromosome to autosomal intensities. (iii) Disproportionate missingness or heterozygosity compared to the cohort-wide fraction of called SNPs that were missing or heterozygous respectively. (iv) Identity by descent (IBD) at 5% or higher, identified by a hidden Markov model (HMM). (v) Identity checks. Prior to full genotyping, individuals were genotyped by Sequenom at a reference panel of 30 SNPs via the Wellcome Trust Sanger Institute (WTSI). Samples with concordance of less than 90% between Sequenom and full genotypes were deemed unknown and thus excluded. (vi) Batch effects. The cohort-wide mean of the A and B allele intensities from 10000 SNPs on chromosome 22 was computed and outliers were excluded.

In addition, exclusions regarding SNPs included minor allele frequency less than 0.01, violation of Hardy-Weinberg equilibrium (HWE) at < 1e-20 and genotyping missingness of 2% or more at the SNPs that reach the maximum call probability threshold.

Statistical analysis of data from PennCNV

The combination of LRR and BAF values was used to estimate *KLK3* deletion frequencies in two large samples by PennCNV.

The two essential metrics to infer copy number states are the LRR (standardised to have a mean=0) and the BAF (ranges from 0 to 1).[21] In an Illumina data analysis framework, X and Y values are derived for each marker from raw data in a five-step normalisation procedure implemented in Illumina's software BeadStudio. Normalisation includes, among other steps,

removing outliers, background estimation and scale appropriation.[22] R and theta (θ) values represent polar transformations of the intensities, and from this (and reference cluster locations) the LRR and BAF metrics are calculated.

The R ratio obtained by PennCNV normalises the signal intensity from an individual to that of a reference sample population for each marker in the dataset. The R, or total signal intensity of both alleles at a SNP (R=X+Y) is calculated, after which LRR [$\log_2(R_{subject}/R_{expected})$] can then be computed.[22] The $R_{expected}$ is computed from "canonical genotype clusters" detailed in.[21] The theta value is calculated as $\theta = \arctan(Y/X)/(\pi/2)$, and is meant to measure the relative allelic intensity ratio of alleles A and B.

The transformation of R and θ into LRR and BAF aims to standardise signal level data across SNPs in order to allow ready comparisons while remaining sufficiently close to the raw data to carry relevant information for copy number quantification.[23]

RESULTS

ARCS assays at *KLK3* exon 3

A total of 152 samples showed evidence of *KLK3* deletion at exon 3 in at least one ARCS replicate. Among these deletions, 31 samples had both ARCS duplicates with higher values that 1.96SDs. A total of 18 samples had a failed replicate, and thus were removed from subsequent plate-specific statistical copy number assignment. A full list of samples with *KLK3* deletions, in addition to samples with PCR failures, is provided (Table 1). All of the deletions identified were single copy deletions. On the other hand, 116 samples were in accordance with a potential copy number gain in *KLK3* at exon 3. Supplementary Figure 1 shows results from one of the 1958 cohort DNA plates as an example of CNV calling defined by the 95% confidence interval.

SNP array data (including BAF and LRR): WTCCC datasets. PennCNV

CNV data from the analysis of the 1958BC data on PennCNV indicate the presence of 61 heterozygous deletions encompassing *KLK3* (freq= 2.13%) (Figure 1).

The majority of *KLK3* deletions in this cohort (86.5%) were less than 50 kb in length (median CNV length≈15.5 kb). LRR and BAF plots for one deletion detected by PennCNV are shown in (Supplementary Figure 2). A full list of deletions called in this cohort is provided(Appendix 1).

PennCNV identified three putative homozygous deletions (copy number = 0). Detailed manual inspection of log R Ratios and B allele frequency values were inconsistent with the occurrence of homozygous deletions.

A subset of 44 deletions in *KLK3* from the 1958BC data (≈61%) start at the SNP rs3760722 in the promoter of the gene (Figure 2).

These deletions range in length between 5.2 and 44 kb (mean=12.6 kb; median=6.8 kb; median no. of SNPs=18). Common endpoints were observed for these deletions, at rs6998 (14 deletions), rs2735839 (13 deletions), rs6072 (4 deletions), rs2739472 and rs198972 (3 deletions each) and rs198979 (2 deletions). Other common start points for deletions within this set include the SNPs rs2003783 (3 deletions), rs266881 (3 deletions), rs2271095 (2 deletions). Common start points, end points and deletions are listed in Appendix 1.

PennCNV analysis of the NBS cohort revealed a lower number of deletions compared to the 1958 Cohort (27 deletions, freq= 0.99%) as shown (Figure 3).

Deletions within this set tend to be shorter (mean length≈26.6 Kb, median≈17.9 Kb) than the 1958BC deletions (Table 2).

All of the deletions detected were hemizygous (Appendix 1). As in the 1958BC dataset, 12 deletions identified from the NBS cohort (44.8%) start at rs3760722; whereas 3 deletions

start at each one of the SNPs rs2659051, rs266849 and rs266881 (Figure 3). Common start points, end points and deletions are listed in Appendix 2.

The two-fold difference (2.12% vs 0.99%) observed in the two samples tested is significant ($P<0.05$).

DISCUSSION

To our knowledge, this study is the first report of the frequency of *KLK3* deletions in the general population. We also found a number of *KLK3* deletions with the same breakpoints, and others displaying one common breakpoint. Data from PennCNV estimates the frequency of *KLK3* deletions within the 1958BC at 0.0213, and in the NBS cohort at 0.0099. These differences are significant, suggesting the existence of inter-population variability in the frequency of *KLK3* deletions. An alternative explanation suggesting sample error could be argued from the fact that both estimates are in general agreement with our *in vitro* experiments for *KLK3* deletions, estimated to occur at frequencies between 0.0081 and 0.0398. This frequency is considerably higher than that of rare genetic variants and invites questions about the implications that *KLK3* deletions could have on PCa diagnosis, monitoring and screening based on PSA. Assuming that the frequency of deletions observed in our population samples apply to other populations, one could predict the existence of 1 – 2% of individuals that could potentially have a mistakenly low serum PSA concentration as a consequence of the deletions identified. All analysed samples are of Caucasian ethnicity, which minimises the occurrence of population stratification.

CNV calling from ARCS data for samples from the 1958 cohort was based on a stringent measure: ARCS ratio values outside the confines of ±1.96SDs from the mean ratio of all samples per plate (a plate-specific attribute). One could envisage that the ARCS assays used in this study were adequately robust to capture CNV events in cluster analysis. ARCS is an accurate method –especially at low copy numbers- that avoids assumptions drawn from standard curve estimations, and offers a high-throughput, generic framework applicable to any CNV assay.[12]

In silico analyses presented in this work are informative for the detection of *KLK3* deletions. With more than 1.2 million probes, the Illumina 1.2-M Duo used in genotyping the WTCCC2 control sets is the most useful array for this purpose, as it offers the highest genomic resolution available to determine boundaries of deletions.

All deletions were found in heterozygosis. However, in three instances, PennCNV assigned a zero copy number to three samples pointing to apparent homozygous deletions in *KLK3*. Careful inspection of LRR/BAF profiles of these samples did not confirm the occurrence of these apparent homozygous deletions in our population samples.

Under the terms of data and material access agreements, it was not possible to link *in vitro* and *in silico* findings from this study. It has been suggested that biological confirmation is a necessary condition for validation of copy number events.[14] Our *in vitro* and *in silico* estimates represent considerable agreement and independent biological confirmation of the frequency of *KLK3* deletions in the UK population. Future studies are required to estimate the frequency of *KLK3* deletions in other populations.

Our study also enabled us to stratify *KLK3* deletions by type and size. This approach identified a number of deletions that shared the same SNPs at the putative starting and ending points identified by PennCNV. Although it could be argued that the overall frequency of *KLK3* deletions is relatively large, we have observed that the commonest single deletion observed in several individuals, was present in a frequency of 0.005. This frequency is lower than the threshold (0.01) previously suggested [24] to define common CNVs. The occurrence of what could possibly be the same deletion in several individuals from our UK sample suggests

the possibility of identity by descent and hence of transmission through the generations. However, it is also possible that they represent mutation hotspots, i.e. for recurrent similar deletions, (e.g. similar hemoglobin A, HBA, deletions).[25] Neutral or deleterious effects of the deletions found in our study will need to be determined in future studies.

As arrays tend to underestimate the size of CNVs, CNV breakpoints are better characterised by sequencing data. Arrays are limited in terms of resolution by the number and the genomic distribution of SNPs on the genotyping chip. The first and the last markers showing evidence of copy number in an array may still be within a deletion rather than representing its boundaries. Differences between sequence-detected and array-detected breakpoints can be used to evaluate the resolution and the reproducibility of array data.[20]

The promoter SNP rs3760722 emerges as a plausible putative breakpoint at the start of *KLK3* deletions, as shown by the majority of deletions detected by PennCNV. The 3'UTR SNPs rs6998 and rs2735839 were found to represent the endpoint of numerous deletions as well. Interestingly, the latter SNPs were associated with serum PSA concentrations in several studies. Free serum PSA concentrations were associated with rs6998 [26] and rs2735839 was also associated with PCa risk[4] and serum PSA[27-29] with studies suggesting that the association signal from this SNP is implied within a stronger signal from a coding SNP in exon 4 of the *KLK3* gene; rs17632542.[5, 28] The characterisation of well-defined CNVs in *KLK3* would allow for testing associations between *KLK3* deletions and serum PSA concentrations, and the recognition of putative breakpoints from SNP array data could prove a useful guide for *KLK3* resequencing studies.

No data on serum PSA concentrations or PCa status are available from the 1958BC and NBS. Future studies are required in order to derive direct conclusions about the impact of *KLK3* deletions on the correlation between gene dose effect and serum PSA and PCa status. This will inform about the suitability of the PSA test in individuals carrying *KLK3* deletions.

REFERENCES

1. Lilja H, Ulmert D and Vickers AJ. (2008) Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nat Rev Cancer* 8: 268-278.
2. Lawrence MG, Lai J and Clements JA. (2010) Kallikreins on steroids: structure, function, and hormonal regulation of prostate-specific antigen and the extended kallikrein locus. *Endocr Rev* 31: 407-446.
3. Cramer SD, Chang BL, Rao A, et al. (2003) Association between genetic polymorphisms in the prostate-specific antigen gene promoter and serum prostate-specific antigen levels. *J Natl Cancer Inst* 95: 1044-1053.
4. Eeles RA, Kote-Jarai Z, Giles GG, et al. (2008) Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet* 40: 316-321.
5. Gudmundsson J, Besenbacher S, Sulem P, et al. (2010) Genetic correction of PSA values using sequence variants associated with PSA levels. *Sci Transl.Med* 2: 62ra92.
6. Kote-Jarai Z, Amin Al OA, Leongamornlert D, et al. (2011) Identification of a novel prostate cancer susceptibility variant in the KLK3 gene transcript. *Hum.Genet.* 129: 687-694.
7. Rodriguez S, Al-Ghamdi OA, Burrows K, et al. (2013) Very low PSA concentrations and deletions of the KLK3 gene. *Clin.Chem.* 59: 234-244.
8. Charbonnier F, Raux G, Wang Q, et al. (2000) Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments. *Cancer Res* 60: 2760-2763.
9. Armour JA, Sismani C, Patsalis PC, et al. (2000) Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res* 28: 605-609.
10. Gille JJ, Hogervorst FB, Pals G, et al. (2002) Genomic deletions of MSH2 and MLH1 in colorectal cancer families detected by a novel mutation detection approach. *Br J Cancer* 87: 892-897.
11. Armour JA, Palla R, Zeeuwen PL, et al. (2007) Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Res* 35: e19.
12. Guthrie PA, Gaunt TR, Abdollahi MR, et al. (2011) Amplification ratio control system for copy number variation genotyping. *Nucleic Acids Res* 39: e54.
13. Ceulemans S, van der Ven K and Del-Favero J. (2012) Targeted screening and validation of copy number variations. *Methods Mol Biol* 838: 311-328.
14. Winchester L, Yau C and Ragoussis J. (2009) Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic* 8: 353-366.
15. Wang K, Li WD, Glessner JT, et al. (2010) Large copy-number variations are enriched in cases with moderate to extreme obesity. *Diabetes* 59: 2690-2694.

16. Uyan O, Omur O, Agim ZS, et al. (2013) Genome-wide copy number variation in sporadic amyotrophic lateral sclerosis in the Turkish population: deletion of EPHA3 is a possible protective factor. *PLoS.ONE.* 8: e72381.
17. Atherton K, Fuller E, Shepherd P, et al. (2008) Loss and representativeness in a biomedical survey at age 45 years: 1958 British birth cohort. *J Epidemiol Community Health* 62: 216-223.
18. Power C and Elliott J. (2006) Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* 35: 34-41.
19. Wang K, Li M, Hadley D, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17: 1665-1674.
20. Pinto D, Darvishi K, Shi X, et al. (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 29: 512-520.
21. Peiffer DA, Le JM, Steemers FJ, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 16: 1136-1148.
22. Wang K and Bucan M. (2008) Copy Number Variation Detection via High-Density SNP Genotyping. *CSH Protoc* 2008: pdb.top46.
23. Wang H, Veldink JH, Blauw H, et al. (2009) Markov Models for inferring copy number variations from genotype data on Illumina platforms. *Hum Hered* 68: 1-22.
24. Kuningas M, Estrada K, Hsu YH, et al. (2011) Large common deletions associate with mortality at old age. *Hum.Mol.Genet.* 20: 4290-4296.
25. Galanello R and Cao A. (2011) Gene test review. Alpha-thalassemia. *Genet.Med.* 13: 83-88.
26. Klein RJ, Hallden C, Cronin AM, et al. (2010) Blood biomarker levels to aid discovery of cancer-related single-nucleotide polymorphisms: kallikreins and prostate cancer. *Cancer Prev Res (Phila)* 3: 611-619.
27. Wiklund F, Zheng SL, Sun J, et al. (2009) Association of reported prostate cancer risk alleles with PSA levels among men without a diagnosis of prostate cancer. *Prostate* 69: 419-427.
28. Parikh H, Wang Z, Pettigrew KA, et al. (2011) Fine mapping the KLK3 locus on chromosome 19q13.33 associated with prostate cancer susceptibility and PSA levels. *Hum Genet* 129: 675-685.
29. Eeles RA, Kote-Jarai Z, Al Olama AA, et al. (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* 41: 1116-1121.

Figure legends

Supplementary Figures

Supplementary Figure 1. ARCS copy number analysis.

ARCS ratios are derived from reference and target peak heights which are plotted on the x-axis above, with fluorescence intensity on the y-axis. Samples with ARCS ratios in excess of 2SDs from the mean ARCS ratio per run are classified as deletion candidates (e.g. sample 10E), whereas samples with ratios lower the 2SDs are deemed duplication candidates (e.g. samples 12A and 3C).

Supplementary Figure 2. Typical LRR and BAF profiles of a deletion detected by PennCNV.

PennCNV uses LRR and BAF values to inform an HMM in order to assign copy numbers to deletions and duplications from SNP array data. This figure shows images of a deletion generated by the visualisation function in PennCNV. This script uses the statistical package R to generate an LRR/BAF profile for each sample. LRR plots are on the top panel, while BAF plots are on the bottom. The LRR/BAF values at each marker are represented by a dot. Each profile is segmented into three parts with left and right flanking regions showing normal copy number patterns and the discovered CNV in the middle demarcated between two grey vertical lines. The blue dots represent LRR/BAF values in regions with normal copy number, while the red dots correspond to the called CNV. The chromosomal positions are on the x-axis of each plot. With normal copy number, BAF values cluster near to 0.0, 0.5 and 1.0, whereas in a hemizygous deletion, BAF values cluster near 0.0 or 1.0, but not around 0.5 -as seen in the red dots between the vertical lines in the left image-. As for the LRR signals, these are normalised to 0.0, so that a deletion will show a decrease (typically around -0.5 for hemizygous deletions, but much lower for homozygous deletions).

Print Figures

Figure 1. *KLK3* heterozygous deletions in the 1958BC dataset.

A snapshot of custom tracks generated by the UCSC genome browser. 61 single copy deletions were found in the 1958BC. Each deletion is represented by a track –in black-. (*KLK3* is highlighted by the red box in the bottom panel. The scale bar at the top of the panel represents 100 kb.

Figure 2. *KLK3* deletions with a common start breakpoint in 1958BC.

A total of 44 deletions were found to start at the SNP rs3760722 in the 1958BC dataset. Common endpoints are found at 6 different SNPs (annotated in red beside each group of tracks that share the same end). (*KLK3* is highlighted by the red box in the bottom panel. The scale bar at the top of the panel represents 20 kb.

Figure 3. *KLK3* deletions in NBS A) Length. B) Common start breakpoints

A) A total of 27 deletions were identified from the NBS dataset, with each deletion represented as a track –in black-. Several recurrent breakpoints could be identified visually. For example, the *KLK3* promoter SNP rs3760722 marks the start of 12 deletions (*KLK3* is highlighted by the red box in the lower panel of the figure, generated by the UCSC genome browser. The scale bar at the top of the panel represents 100 kb.

B) 12 deletions in the NBS dataset started at rs37607222. Four deletions were found to share putative start and end breakpoints. (*KLK3* is highlighted by the red box in the lower panel of the figure, generated by the UCSC genome browser. The scale bar at the top of the panel represents 20 kb.

Table 1. A list of samples where *KLK3* exon 3 was found deleted by ARCS.

Deletions		**Failed ARCS**	
2 reps	**1 rep**	**2 reps**	**1 rep**
8H/12F/12H	9A/12G		
2H	2E/5G/12D		
8A	8H		
8A/9D	6H/9C		
2E	2A/3C/5E/7A		
11D	4C/7E/11A		
12F/12G	11F/12H		
	4G/6D/7B		5(A-H) /8A
2E/5F/12H			
	8F/12H		
	7E		
	2F/2G/6B		
11B	8E/10G		
	2D		12F/12G/12H
	9B/10C		
8G	2H/10D/11B		
	3B/3E/8B/9C		
	4G/5G/6F/7G		2H
	9H/10H/11H/12E		
	2D/2G/11E		
	3E/10H/11G	2G	
	6F/10H/11G/11H		
	6E/8F/12E		
	2H/4H/5G/7G		
4G/12H	4D/6E/6F/12E		
2A	2F/11H/12H		
	3C/4D/4H/10C/11B		
	5B/10C/10F		12F/12G/12H
3F/12H	12E/12G		
9E	7D/10A/10G		3E
10E			
2F	8C/8G/12B		
	4E		
	3H/8A/8E		
	2A/4E/5G/9B/12A/12B		
2A	3A/8D/8H/9F		
4H			
	7H/11A/11H		
	11H/12E		
	2H/3C/8G		
8A/9D	7A/7F/9A/9B/10B/11H/12A		
12F	9G		
	2G/2H/5F/11A		
2A/12C	12D		
31	**121**	1	17

152 samples have shown evidence for a deletion in KLK3 exon 3 by at least one replicate in ARCS, including 31 samples where both replicates indicated a deletion. 18 samples were PCR failures.

Table 2. Results from the 1958BC and NBS datasets – PennCNV

WTCCC2 dataset	Count (Freq)	Median SNPs	ROH Length (Kb)			
			Shortest	Longest	Mean	Median
1958 BC	61 (2.13%)	18	5.2	81.2	17.4	6.8
NBS	27 (0.99%)	27	2.3	72.4	25.5	17.9